PATENT
Attorney Docket No. 0026-0016

UNITED STATES PATENT APPLICATION

OF

LUIS GRAVANO and MONIKA H. HENZINGER

FOR

SYSTEMS AND METHODS FOR USING ANCHOR TEXT AS PARALLEL CORPORA FOR CROSS-LANGUAGE INFORMATION RETRIEVAL

[0001] SYSTEMS AND METHODS FOR USING ANCHOR TEXT AS PARALLEL CORPORA FOR CROSS-LANGUAGE INFORMATION RETRIEVAL

[0002] BACKGROUND OF THE INVENTION

[0003] Field of the Invention

[0004] The present invention relates generally to information retrieval systems and, more particularly, to systems and methods for translating search queries for cross-language information retrieval.

[0005] Description of Related Art

[0006] Many users of a hypertext medium, such as the World Wide Web ("web"), can read documents in more than one language. Consider, for example, a query in English from a user that can read English and Spanish. A conventional technique for identifying documents in Spanish for this English query involves translating the query to Spanish and then processing the translated query to identify matching Spanish documents.

[0007] Query terms are inherently ambiguous. Therefore, translating them is challenging. Some conventional approaches use a bilingual dictionary to perform query translations. It has been found, however, that using a bilingual dictionary results in noisy translations. The noisy translations may be due to many factors. For example, a translation may result in extraneous terms being added to the query because a dictionary entry may list several senses for a term. In other words, each term may have one or more possible translations in the dictionary. Also, general dictionaries often do not include technical terminology. This makes translation of technical query terms difficult.

[0008] Other conventional approaches rely either on "parallel corpora" (i.e., collections of documents in which each of the documents appears in two different languages) or "co-occurrence statistics" of terms in documents in the target language to which the query is being translated to translate query terms. A problem with the parallel corpora approach is that such corpora are rare and building them is prohibitively expensive.

[0009] As a result, there exists a need for mechanisms that translate queries to facilitate cross-language information retrieval.

[0010] SUMMARY OF THE INVENTION

[0011] Systems and methods consistent with the present invention address this and other needs by providing mechanisms for translating search queries that exploit anchor text in one language that refer to documents in another language to produce good quality, less noisy query translations.

[0012] In accordance with the principles of the invention as embodied and broadly described herein, a system performs cross-language query translations. The system receives a search query that includes terms in a first language and determines possible translations of the terms of the search query into a second language. The system also locates documents in the first language that contain references that match the terms of the search query and identify documents in the second language. The system then disambiguates among the possible translations of the terms of the search query using the identified documents to identify one of the possible translations as a likely translation of the search query into the second language.

[0013] In another implementation consistent with the present invention, a method for performing cross-language document retrieval is provided. The method includes receiving a search query that includes one or more terms in a first language; performing a search of documents in the first language to locate one or more of the first language documents that contain anchor text that matches the search query and identifies one or more documents in a second language; determining possible translations of the terms of the search query into the second language; using the identified second language documents as parallel corpora for disambiguation among the possible translations of the terms of the search query; identifying one of the possible translations as a correct translation of the search query based on the disambiguation; and performing a search of second language documents using the correct translation of the search query.

[0014] In yet another implementation consistent with the present invention, a system for performing cross-language query translations is provided. The system receives a search query that includes terms in a first language and determines possible translations of the terms of the search query into a second language. The system also locates documents in the first language that contain references that match the terms of the search query and refer to other documents in the first language and identify documents in the second language that contain references to the other documents. The system then disambiguates among the possible translations of the terms of the search query using the identified documents to identify one of the possible translations as a likely translation of the search query.

[0015] In a further implementation consistent with the present invention, a method for performing cross-language query translation is provided. The method includes receiving a search query that includes terms in a first language; determining possible translations of the terms of the search query into a second language; locating documents in the first language that match the terms of the search query; identifying documents in the second language that contain references to the first language documents; and disambiguating among the possible translations of the terms of the search query using the second language documents to identify one of the possible translations as a likely translation of the search query.

[0016] BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

[0018] Fig. 1 is a diagram of an exemplary network in which systems and methods consistent with the present invention may be implemented;

[0019] Fig. 2 is an exemplary diagram of a server of Fig. 1 in an implementation consistent with the present invention;

[0020] Fig. 3 is an exemplary functional diagram of a query translation portion of the server of Fig. 2 according to an implementation consistent with the present invention;

[0021] Fig. 4 is a diagram illustrating relations between an exemplary set of web documents that may be stored in the database of Fig. 3;

[0022] Fig. 5 is a flowchart of exemplary processing for translating a search query in accordance with an implementation consistent with the present invention;

[0023] Fig. 6 is a diagram of an exemplary graphical user interface that may be presented to the user to facilitate the providing of search information;

[0024] Fig. 7 is a flowchart of exemplary processing for performing query translation in accordance with an alternate implementation consistent with the present invention; and

[0025] Fig. 8 is a flowchart of exemplary processing for performing query translation in accordance with another implementation consistent with the present invention.

[0026] DETAILED DESCRIPTION

[0027] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents.

[0028] Systems and methods consistent with the present invention translate search queries to facilitate information retrieval in different languages. The systems and methods use anchors in one language that link to documents in another language to produce good quality, less noisy translations.

[0029] EXEMPLARY NETWORK CONFIGURATION

[0030] Fig. 1 is an exemplary diagram of a network 100 in which systems and methods consistent with the present invention may be implemented. The network 100 may include

multiple clients 110 connected to multiple servers 120-130 via a network 140. The network 140 may include a local area network (LAN), a wide area network (WAN), a telephone network, such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, a similar or dissimilar network, or a combination of networks. Two clients 110 and three servers 120-130 have been illustrated as connected to network 140 in Fig. 1 for simplicity. In practice, there may be more or fewer clients and servers. Also, in some instances, a client may perform the functions of a server and a server may perform the functions of a client.

[0031] The clients 110 may include devices, such as wireless telephones, personal computers, personal digital assistants (PDAs), laptops, or other types of communication devices, threads or processes running on these devices, and/or objects executable by these devices. The servers 120-130 may include server devices, threads, and/or objects that operate upon, search, or maintain documents in a manner consistent with the present invention. The clients 110 and servers 120-130 may connect to the network 140 via wired, wireless, or optical connections.

[0032] In an implementation consistent with the present invention, the server 120 may include a search engine usable by the clients 110. The servers 130 may store documents, such as web documents or web pages, accessible by the clients 110 and the server 120.

[0033] EXEMPLARY SERVER ARCHITECTURE

[0034] Fig. 2 is an exemplary diagram of the server 120 in an implementation consistent with the present invention. The server 120 may include a bus 210, a processor 220, a main memory 230, a read only memory (ROM) 240, a storage device 250, one or more input devices 260, one

or more output devices 270, and a communication interface 280. The bus 210 may include one or more conductors that permit communication among the components of the server 120.

[0035] The processor 220 may include any type of conventional processor or microprocessor that interprets and executes instructions. The main memory 230 may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by the processor 220. The ROM 240 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by the processor 220. The storage device 250 may include a magnetic and/or optical recording medium and its corresponding drive.

[0036] The input devices 260 may include one or more conventional mechanisms that permit an operator to input information to the server 120, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. The output devices 270 may include one or more conventional mechanisms that output information to the operator, including a display, a printer, a speaker, etc. The communication interface 280 may include any transceiver-like mechanism that enables the server 120 to communicate with other devices and/or systems. For example, the communication interface 280 may include mechanisms for communicating with another device or system via a network, such as network 140.

[0037] As will be described in detail below, the server 120, consistent with the present invention, perform certain searching-related operations. The server 120 may perform these operations in response to processor 220 executing software instructions contained in a computer-

readable medium, such as memory 230. A computer-readable medium may be defined as one or more memory devices and/or carrier waves.

[0038] The software instructions may be read into memory 230 from another computer-readable medium, such as the data storage device 250, or from another device via the communication interface 280. The software instructions contained in memory 230 causes processor 220 to perform processes that will be described later. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes consistent with the present invention. Thus, the present invention is not limited to any specific combination of hardware circuitry and software.

[0039] QUERY TRANSLATION MECHANISMS

[0040] Fig. 3 is an exemplary functional diagram of a portion of the server 120 according to an implementation consistent with the present invention. This portion of the server 120 may be implemented in software, hardware, or a combination of software and hardware.

[0041] The portion of the server 120 shown in Fig. 3 includes a database 310, a search engine 320, a dictionary 330, and a query translation engine 340. The database 310 may store copies of web documents stored by other servers 130 in the network 140 and/or a tree or directed graph of linked web documents. The documents and directed graph may be stored in the database 310 by a spider program that "crawls" web documents on network 140 based on their hyperlinks. As a result, the spider program may generate a tree or directed graph of linked web documents. The crawled documents may be stored in the database as an inverted index in which each term in the database 310 is associated with all the crawled documents that contain that term.

[0042] Fig. 4 is a diagram illustrating relations between an exemplary set of web documents that may be stored in the database 310. Documents 410-430 contain links to each other. In the nomenclature of the popular Hyper-Text Markup Language (HTML) standard, hyperlinks to other documents are specified using an HTML structure called an anchor. An anchor specifies the Uniform Resource Locator (URL) of the document being linked. Typically, browsers display anchors as text distinct from the main text using, for example, underlining or different colors. A user, by selecting the anchor, causes the browser to retrieve the web document specified by the URL.

[0043] In Fig. 4, documents 410-430 contain various links. In particular, document 410 contains an anchor 412 that corresponds to a forward link to document 420 and an anchor 414 that corresponds to a forward link to document 430. Document 420 contains an anchor 422 that corresponds to a forward link to document 430. Document 430 contains an anchor 432 that corresponds to a forward link to document 410.

[0044] Returning to Fig. 3, the search engine 320 may include logic that identifies one or more documents or a list of documents in response to a search query that includes one or more search terms. For example, the search engine 320 may receive a search query from a user and respond by returning relevant information or a list of relevant information to the user. Typically, users may ask the server 120 to locate web documents relating to a particular topic that are stored at other devices or systems connected to network 140 or another network. The search engine 320 may access the database 310 to compare the terms in the search query to the documents in the database 310.

[0045] The dictionary 330 may include one or more bilingual machine-readable dictionaries. The dictionary 330 may provide information to facilitate translations between multiple languages. The information in the dictionary 330 may be used by the query translation engine 340 to translate terms in a search query.

[0046] The query translation engine 340 may include logic that translates the terms of a search query using information from the dictionary 330 and the search engine 320. For example, the query translation engine 340 may identify potential translations for the terms of a search query based on the information in the dictionary 330. The query translation engine 340 may then disambiguate among the potential translations based on text from documents identified by the search engine 320, as described below.

[0047] EXEMPLARY PROCESSING FOR QUERY TRANSLATION

[0048] Fig. 5 is a flowchart of exemplary processing for performing query translation in accordance with an implementation consistent with the present invention. Processing may begin with a user accessing a server, such as server 120 (Fig. 1), using, for example, web browser software on a client, such as client 110. The user may then provide a query that includes one or more search terms to the search engine 320 (Fig. 3) maintained by the server 120 (act 510).

[0049] Assume, for purposes of this example, that the user provides search terms in a first language and desires documents in a second language. To facilitate the providing of information for a search, the server 120 may provide a graphical user interface (GUI) to the user. Fig. 6 is an exemplary diagram of a GUI 600 consistent with the present invention. The GUI 600 may prompt the user to enter one or more search terms/words to include (box 610) or exclude (box

620) in the search results. The GUI 600 may also prompt the user to identify the language(s) in which the results will be presented (box 630).

[0050] Returning to Fig. 5, the search engine 320 may perform a search using the terms of the query in the first language (act 520). In this case, the search engine 320 looks for documents in the first language that contain anchor text that matches the search query and refers to a document in the second language (act 530). When determining whether there is a match between the terms of the search query and the anchor text, the search engine 320 may consider not only the text making up the anchor, but also surrounding text, such as the text in the paragraph containing the anchor. The search engine 320 may then identify the documents in the second language that are referenced by the anchor text in the first documents (act 540) and provides these documents to the query translation engine 340.

[0051] Meanwhile, the query translation engine 340 may perform an initial translation on the terms of the search query. For example, the query translation engine 340 may use the dictionary 330 to identify potential translations for terms in the query (act 550). A dictionary entry may have several senses for a term, however, leading to several possible translations.

[0052] To disambiguate among the potential translations, the query translation engine 340 may use conventional parallel corpora disambiguation techniques, such as the techniques identified in Y. Yang et al., "Translingual Information Retrieval: Learning from Bilingual Corpora," Artificial Intelligence Journal special issue: Best of IJCAI-97, 1998, pp. 323-345, and L. Ballesteros et al., "Resolving Ambiguity for Cross-Language Retrieval," Proceedings of ACM SIGIR, 1998, pp. 64-71, which are incorporated herein by reference. According to an implementation consistent

with the present invention, however, the query translation engine 340 uses the text from the documents in the second language that were identified by the search engine 320 as the parallel corpora (act 560). Because these documents possibly contain text related to the original search query, the translations produced by the query translation engine 340 are of good quality and less noisy.

[0053] The query translation engine 340 may then output the translated query (in the second language) (act 570). The search engine 320, or another search engine, may identify documents in the second language that correspond to the translated query and present the documents to the user.

[0054] EXAMPLE

[0055] Assume that a user provides a search query to the server 120 in Spanish, but desires documents to be returned in English. Further, assume that the user desires documents relating to "banks interest." In this case, the query provided by the user may include the terms "bancos" and "interés." To facilitate English-language document retrieval, the server 120 may translate the Spanish query to English.

[0056] The query translation engine 340 may perform an initial translation of the terms of the query using, for example, the dictionary 330. In this case, the query translation engine 340 finds that each of the terms of the query has more than one possible translation. For example, the Spanish word "bancos" could be translated as "banks" or "benches" (among other possibilities) in English. The Spanish word "interés" could be translated as "interest" or "concern" (among other

possibilities) in English. The query translation engine 340 disambiguates among the possible translations using documents identified by the search engine 320.

[0057] The search engine 320 performs a search using the original Spanish query (i.e., "bancos interés") to identify Spanish-language documents that include anchors that contain all of the query terms and point to English-language documents. The search engine 320 provides the English-language documents that are pointed to by the anchors to the query translation engine 340.

[0058] The query translation engine 340 analyzes the text of the English-language documents to, for example, compute the frequency of co-occurrence of the various translation possibilities. Specifically, the query translation engine 340 determines how often the word "banks" occurs with "interest," "banks" occurs with "concern," "benches" occurs with "interest," and "benches" occurs with "concern." Presumably, the query translation engine 340 would determine that "banks" and "interest" are the most frequent combination and use these terms as the correct translation for the Spanish query "bancos interés."

[0059] ALTERNATE IMPLEMENTATIONS

[0060] Fig. 7 is a flowchart of exemplary processing for performing query translation in accordance with an alternate implementation consistent with the present invention. Processing may begin with a user accessing a server, such as server 120 (Fig. 1), using, for example, web browser software on a client, such as client 110. The user may then provide a query that includes one or more search terms to the search engine 320 (Fig. 3) maintained by the server 120 (act 710).

[0061] Assume, for purposes of this example, that the user provides search terms in a first language and desires documents in a second language. To facilitate the providing of information for a search, the server 120 may provide a GUI to the user, such as the one illustrated in Fig. 6.

[0062] The search engine 320 may perform a search using the terms of the query in the first language (act 720). In this case, the search engine 320 looks for documents in the first language that contain anchor text that matches the search query and references another document in the first language (act 730). When determining whether there is a match between the terms of the search query and the anchor text, the search engine 320 may consider not only the text making up the anchor, but also surrounding text, such as the text in the paragraph containing the anchor.

[0063] The search engine 320 may then identify documents in the second language that contain anchor text that refers to the referenced documents in the first language (act 740). The search engine 320 may provide these documents to the query translation engine 340.

[0064] Meanwhile, the query translation engine 340 may perform an initial translation on the terms of the search query. For example, the query translation engine 340 may use the dictionary 330 to identify potential translations for terms in the query (act 750). A dictionary entry may have several senses for a term, however, leading to several possible translations.

[0065] To disambiguate among the potential translations, the query translation engine 340 may use conventional parallel corpora disambiguation techniques, such as the ones described above. According to an implementation consistent with the present invention, however, the query translation engine 340 uses the text from the documents in the second language that were identified by the search engine 320 as the parallel corpora (act 760). The text used by the query

translation engine 340 may include the anchor, text surrounding the anchor, or the entire text of the documents. Because these documents possibly contain text related to the original search query, the translations produced by the query translation engine 340 are of good quality and less noisy.

[0066] The query translation engine 340 may then output the translated query (in the second language) (act 770). The search engine 320, or another search engine, may identify documents in the second language that correspond to the translated query and present the documents to the user.

[0067] Fig. 8 is a flowchart of exemplary processing for performing query translation in accordance with another implementation consistent with the present invention. Processing may begin with a user accessing a server, such as server 120 (Fig. 1), using, for example, web browser software on a client, such as client 110. The user may then provide a query that includes one or more search terms to the search engine 320 (Fig. 3) maintained by the server 120 (act 810).

[0068] Assume, for purposes of this example, that the user provides search terms in a first language and desires documents in a second language. To facilitate the providing of information for a search, the server 120 may provide a GUI to the user, such as the one illustrated in Fig. 6.

[0069] The search engine 320 may perform a search using the terms of the query in the first language (act 820). In this case, the search engine 320 looks for documents in the first language that contain text that matches the search query (act 830). For this implementation, the search engine 320 may match the terms of the query to any text in the documents. The search engine 320 may then identify documents in the second language that contain anchor text that refers to

the documents in the first language (act 840). The search engine 320 may provide these documents to the query translation engine 340.

[0070] Meanwhile, the query translation engine 340 may perform an initial translation on the terms of the search query. For example, the query translation engine 340 may use the dictionary 330 to identify potential translations for terms in the query (act 850). A dictionary entry may have several senses for a term, however, leading to several possible translations.

[0071] To disambiguate among the potential translations, the query translation engine 340 may use conventional parallel corpora disambiguation techniques, such as the ones described above. According to an implementation consistent with the present invention, however, the query translation engine 340 uses the text from the documents in the second language that were identified by the search engine 320 as the parallel corpora (act 860). The text used by the query translation engine 340 may include the anchor, text surrounding the anchor, or the entire text of the documents. Because these documents possibly contain text related to the original search query, the translations produced by the query translation engine 340 are of good quality and less noisy.

[0072] The query translation engine 340 may then output the translated query (in the second language) (act 870). The search engine 320, or another search engine, may identify documents in the second language that correspond to the translated query and present the documents to the user.

[0073] CONCLUSION

[0074] Systems and methods consistent with the present invention provide good quality, less noisy search query translations by exploiting anchors in one language that point to documents in another language.

[0075] The foregoing description of preferred embodiments of the present invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. For example, while a series of acts has been described with regard to Fig. 5, the order of the acts may be modified in other implementations consistent with the present invention. Moreover, non-dependent acts may be performed in parallel.

[0076] Also, it has been described that the database 310, the search engine 320, and the query translation engine 340 are located on the same server 120. In other implementations consistent with the present invention, the database 310, the search engine 320, and/or the query translation engine 340 are located on different systems or devices.

[0077] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used. The scope of the invention is defined by the claims and their equivalents.